Animal (2007), **1:8**, pp 1179–1187 © The Animal Consortium 2007
doi: 10.1017/S1751731107000535

Aggregation of measures to produce an overall assessment of animal welfare. Part 1: a review of existing methods

R. Botreau[1,2†], M. Bonde[3], A. Butterworth[4], P. Perny[5], M. B. M. Bracke[6], J. Capdeville[2] and I. Veissier[1]

[1]*INRA, UR1213 Herbivores, Site de Theix, F-63122 Saint-Genès-Champanelle, France;* [2]*Institut de l'Elevage, BP18, Castanet Tolosan F-31321, France;* [3]*Department of Animal Health, Welfare and Nutrition, Danish Institute of Agricultural Sciences, POB 50, Tjele DK-8830, Denmark;* [4]*University of Bristol Clinical Veterinary Science, Langford UK-BS40 5DU, UK;* [5]*Laboratoire Informatique de Paris 6, Université Pierre et Marie Curie, 8 rue du Capitaine Scott, Paris F-75015, France;* [6]*Animal Sciences Group, Wageningen University and Research Centre, POB 65, Lelystad NL-8200 AB, The Netherlands*

(Received 10 January 2007; Accepted 6 June 2007)

Several systems have been proposed for the overall assessment of animal welfare at the farm level for the purpose of advising farmers or assisting public decision-making. They are generally based on several measures compounded into a single evaluation, using different rules to assemble the information. Here we discuss the different methods used to aggregate welfare measures and their applicability to certification schemes involving welfare. Data obtained on a farm can be (i) analysed by an expert who draws an overall conclusion; (ii) compared with minimal requirements set for each measure; (iii) converted into ranks, which are then summed; or (iv) converted into values or scores compounded in a weighted sum (e.g. TGI35L) or using ad hoc *rules. Existing methods used at present (at least when used exclusively) may be insufficiently sensitive or not routinely applicable, or may not reflect the multidimensional nature of welfare and the relative importance of various welfare measures. It is concluded that different methods may be used at different stages of the construction of an overall assessment of animal welfare, depending on the constraints imposed on the aggregation process.*

Keywords: animal welfare, assessment, livestock, methodology

Introduction

During the last 30 years, extensive experimental work has been carried out to collect information and to inform the societal debate on animal welfare. The aim of much of this study has been to describe the impact of living conditions on animals and to gain a better understanding of their needs, preferences or aversions (Rushen, 1986; Dawkins, 1990). These studies have resulted in recommendations to producers or policy makers on how animals should be kept for farming or other purposes (see for instance the reports from the Scientific Committee on Animal Health and Animal Welfare of the European Commission, 2007). More recently, researchers have attempted to assess the actual welfare of farm animals on farms or at slaughter to provide information on the influences on animal welfare of their living (or slaughter) conditions. Such information is considered by many to be essential in the context of certification schemes based, at least in part, on animal welfare (e.g. Bartussek, 1999; Main *et al.*, 2001). These schemes can help improve animal welfare through market-driven strategies (Bock and Van Leeuwen, 2005) and may require a formal model for the overall evaluation of animal welfare (Fraser, 2003).

Welfare is a multidimensional concept: it embraces absence of suffering, high levels of biological functioning – including absence of diseases – and the potential for animals to have 'positive experience' (Fraser, 1993). Each of these principles subsumes several criteria, e.g. absence of suffering comprises absence of prolonged pain, hunger, thirst, fear, discomfort and distress. Some single measures have been proposed to provide a broad assessment of animal welfare including corticosteroids (Barnett and Hemsworth, 1990) (acute phase proteins) (Hurnik, 1990) and longevity (Geers *et al.*, 2003). However, none of these single measures attempts to cover all the dimensions of welfare. For instance, an animal can be diseased with no impairment of its corticotropic axis activity and vice versa.

† E-mail: rbotrea@clermont.inra.fr

https://doi.org/10.1017/S1751731107000535 Published online by Cambridge University Press

Acute phase proteins seem specific to tissue damage (due to disease or aggression between animals) and may not be related to psychological stress (Piñeiro *et al*., 2005). Longevity depends at least in part on the occurrence of diseases and diseases are rather independent of the performance of natural behaviour, which is considered essential to animal welfare (Farm Animal Welfare Council (FAWC), 1992). Hence, it appears likely that several measures are necessary to obtain a comprehensive view of any particular animal's welfare (Dawkins, 1980; Friend, 1980; Webster, 1997; Rutter, 1998).

Sets of measures, made on either the farm or housing environment, or on the animals themselves, have been defined for the purpose of (i) advising farmers on how to improve the welfare of their animals (Sørensen *et al*., 2001); (ii) checking compliance with legislative requirements – for example, as carried out in Sweden in association with the banning of battery cages (Keeling and Svedberg, 1999); (iii) implementing welfare certification schemes (e.g. Freedom Food Scheme: Main *et al*., 2001); or (iv) comparing systems to refine legislation (Bracke *et al*., 2002; for a review see Main *et al*., 2003). Most of these goals require combining measures to form an overall assessment, either to fully describe a situation (e.g. comparing systems) or to make an absolute assessment on it (e.g. checking compliance with standards).

Several types of evaluation models – descriptive, normative and prescriptive – can be distinguished and they play different roles in a decision process based on several measures (Bell *et al*., 1988; Roy, 1993). Descriptive models are use to describe a pre-existing situation that is stable and independent of any observation. For example, multivariate statistics are often used to describe populations and compound variables (i.e. principal components) are calculated to summarise the variability between observations. This approach provides the ability to describe and compare observed situations (as in Veissier *et al*., 2004).

Normative models tell us how things should be or how people should act, and aim at providing evaluation procedures to check the adequacy of observed behaviours relative to pre-defined norms. This is, for example, the case for many formal labelling processes (i.e. certification, hotel/restaurant rating, etc.) and could also be the case for animal welfare evaluation. As observed by French (1984), '… normative analyses tell us how we should behave in particular circumstances, whereas descriptive theories conjecture how things are behaving'.

Finally, prescriptive models are used to help people make better decisions and to improve their activity. The prescriptive approach does not assume any pre-existing situation that should be described but rather aims to gather and organise relevant information so as to facilitate the construction of recommendations to reach a goal. These distinctions in the nature of the model are essential, because they might change the perspective in the construction of the model and the way it should be evaluated or validated. As noted by Bell *et al*. (1988):

> descriptive models are evaluated by their empirical validity, that is the extent to which they correspond to observed choices. Normative models are evaluated by their theoretical adequacy, that is, the degree to which they provide acceptable idealizations or rational choices. Prescriptive models are evaluated by their pragmatic value, that is their ability to help people make better decision.

We do not consider that the evaluation of animal welfare should be seen as a simple descriptive problem because there is no pre-existing universal view of animal welfare that can be measured. More likely, animal welfare assessment is a matter of the definition of norms, evaluation of current practice with respect to these norms and the creation of recommendations (e.g. Fraser, 1995 and 2003). As such, the formal evaluation procedure for assessing animal welfare could be seen as both a normative and a prescriptive model.

Several methods have been proposed to compound the results from welfare measures to make an overall assessment. They range from informal aggregation by experts to the use of precise calculations such as weighted sums of scores obtained for each measure. Here we review the methods that have been proposed for the overall assessment of animal welfare, highlighting their strengths and weaknesses. The relevance of the measures *per se* will not be addressed here as this issue has been largely dealt with elsewhere (e.g. Winckler *et al*., 2003). Our objective is neither to form an overall judgement of the various methods used to produce a welfare assessment nor to rank them. By contrast, we intend to analyse the content and limits of the calculation methods proposed to aggregate welfare measures. This analysis relies on the identification of the theoretical construction of each method. According to the characteristics of each calculation method, we will then analyse in which context(s) a given method could be best used. This should in turn help in choosing the best methods for a specific use, e.g. the comparison of animal units in benchmarking or to provide policy makers with decision tools. More specifically, we will evaluate the potential of each aggregation method for use in the context of certification schemes for animal welfare. For this application, we need an aggregation method that (i) can be easily explained to stakeholders (producers, consumers, etc.); (ii) can be used routinely on large numbers of animal units (farms, slaughter-plants, zoos, etc.) where the composition of these sets varies in time; and (iii) encourages producers to improve animal welfare. The method must therefore be repeatable – so that both farmers and end-users trust the results – and precise – in order to be able to monitor improvements or on the contrary decrements in welfare – while lending greatest importance to the more serious welfare problems – so that major problems can be given first priority and be cured first.

https://doi.org/10.1017/S1751731107000535 Published online by Cambridge University Press

Non-formal aggregation of several welfare measures

When several measures are used to make an overall assessment, the aggregation may not be explicit. This is the case when one expert (or a group of experts) is asked to form an opinion on an animal unit (e.g. a farm or a slaughterhouse) based on a number of observations related to animal welfare. Such non-explicit aggregation of measures is currently used to advise farmers. For instance, Sørensen *et al.* (2001) developed a system for an 'ethical account' of dairy and pig farms, and Hegelund *et al.* (2003) developed a similar system for egg production in organic farms. They take into account animal-based measures (behaviour, health) and environment-based measures (system, management). Welfare problems and their potential causes are identified and a strategy to improve the welfare status of the animals is proposed to the farmer.

Informal appraisal of measures can also be used in certification or authorisation processes. This is the case for systems developed by the Swedish Board of Agriculture to approve or reject new housing systems or equipment for laying hens (Algers *et al.*, 1995; Ekstrand *et al.*, 1997; quoted by Johnsen *et al.*, 2001) or by the Federal Veterinary Office from Switzerland to authorise mass-produced housing systems and equipment designed for farm animals (Wechsler, 2001, 2003 and 2005). Data on production, health, mortality and behaviour of the animals are collected in experimental investigations and on-farm inspections. These are then brought together in a detailed report from which experts are asked to form an opinion on new equipment.

This method seems simple since no mathematical tools are necessary.

However, the rationale behind the process of aggregation is generally not transparent. The opinion is the result of the reasoning of one expert (or several experts), and experts may have different views on the interpretation of measures. For instance, because of their different backgrounds veterinarians may lend more importance to health, whereas ethologists may emphasise behaviour, and so these methods are prone to variability between people.

In addition, animal welfare assessment is often based on a large set of data. As mentioned by Lacroix *et al.* (1997), humans may become saturated by large amounts of detailed information, resulting in some information being discounted when non-formal aggregation of information is used.

Finally, for each evaluation to be performed on a given animal unit, these methods require the opinion and input of 'experts'. Consequently, informal aggregation is not suitable for routine assessments on large numbers of animal units (farms, slaughterhouses) by inexpert persons.

Even so, informal appraisal by experts seems highly appropriate for advising policy makers. In Switzerland, such evaluation has proved efficient for licensing housing systems and equipment. Similarly, the reports from the European Food Safety Agency can be viewed as a summary of expert opinion on farming systems on the basis of several indicators, generally listed at the beginning of the report (e.g. the report on the welfare of calves; Algers *et al.*, 2006).

Definition of minimal requirements for measures

Benchmarking systems are based on minimum requirements that a farm, a housing system or equipment have to meet. A series of measures are taken and a threshold is set for each measure. This system is the basis of the Freedom Food Scheme, which was set up as a farm assurance and food-labelling scheme by the Royal Society for the Prevention of Cruelty to Animals (RSPCA) in 1994. The measures considered are essentially based on resources, management and health records (health plan, diseases, etc.; Main *et al.*, 2001). The Freedom Food scheme has been adapted to North American farms through the certification scheme *'Certified Humane Raised and Handled'* developed by the association for Humane Farm Animal Care (HFAC, 2003). Similarly, Von Borell *et al.* (2001) proposed a method based on hazard analysis and critical control points (HACCP) for pig housing. These authors consider the impact of different housing designs on welfare, health, management and the environment. Key threats to welfare (critical control points) are identified, and adequate controls are then determined for these points. This method may be regarded as a checklist with the same number of points as the critical control points determined in the HACCP procedure.

The methods based on comparison with a list of minimal requirements are straightforward and can be easily explained to laypersons and, unlike a non-formal aggregation, they can be standardised and suit the aim of assuring compliance with pre-set standards. As a consequence, they are perfectly adapted to be implemented in certification schemes based on the fulfilment of a list of specifications, leading to an *'accept/reject'* answer. However, they present some characteristics specific to this aim that may prove to be drawbacks if such methods are to be applied to achieving other objectives.

First, when applied strictly, they are 'all-or-nothing' processes. If the value obtained by a farm for one measure is below the pre-set threshold, then that farm will be considered below standard – and in a certification process, this farm may be rejected. Because of this limitation, these methods may be 'over-reactive': on a farm classified as 'good' a deterioration in only one aspect can result in exclusion from the certification scheme (e.g. a disease outbreak, even if all other parameters remain the same); no distinction is made between a farm that fails on only one aspect and a farm that fails on many aspects. However, some flexibility can be introduced. For instance, in the Freedom Food scheme, when a requirement is not met, the farmer is given 8 weeks to make corrective measures (RSPCA, undated). Hence, problems on a good farm that

https://doi.org/10.1017/S1751731107000535 Published online by Cambridge University Press

fails on one aspect may be corrected within an agreed time, whereas this would be difficult to achieve on a farm that failed on too many (or deep seated) aspects. Another way to make the method more flexible is to allow a certain proportion of non-compliances. A more mathematically elaborate method would be to define reference profiles (of sets of welfare measures or dimensions) and compare observations with these profiles according to majority rules (Bouyssou *et al.*, 2000, p. 226). In this case, the results could take the form of the assignment of a farm to a class corresponding to a certain level of welfare, such as low, moderate or high. This method has not yet been applied in the context of animal welfare.

Second, in these methods, all measures have the same impact, and non-compliance for one measure constitutes a 'veto'. However, some elements, such as painful disease states (e.g. severe lameness), may induce more suffering than others (e.g. lack of positive experiences). Some schemes, such as the EUREPGAP certification standards developed by retailers and their global suppliers (EUREPGAP, nd), include welfare requirements classified as major requirements, minor requirements or recommendations. It is likely that a less-strict compliance is requested for recommendations than for minor requirements and in turn for minor requirements than for major ones.

In conclusion, comparison with a list of minimal requirements (inspection) could be used as part of certification schemes. More-subtle results are obtained when a certain degree of flexibility is introduced and the relative contribution of the different measures for animal welfare is taken into account.

Sum (or mean) of ranks

Whay *et al.* (2003) compared dairy calves' welfare on 45 farms on the basis of 19 measures corresponding to respiratory health, nutrition and general appearance. For each measure the farms were ranked from best (rank 1) to worst (rank 45). Each farm was thus described by 19 partial ranks, and an overall rank was then calculated as the mean of the 19 partial ranks. The farms were sorted according to their overall ranks. Mathematically, this corresponds to the Borda method, i.e. a sum of partial ranks (Bouyssou *et al.*, 2000, p. 129).

In a similar vein, the Bristol Welfare Assurance Programme proposed to rank farms according to the quintile in which the farm is situated among a given population: the farm scores 0 when in the best quintile for a given measure and 4 when in the worst quintile (Webster, 2005). Then the sum of the partial scores obtained by the farm on the various measures is calculated. This method can be assimilated to a sum of ranks since the score obtained by a farm depends on how this farm performs in comparison with other farms, i.e. in this method the partial scores are simplified expressions of partial ranks. An even greater simplification of ranks is produced when farms are simply compared with the average value obtained over the population of the inspected farms, as in Huxley *et al.* (2004) for organic dairy herds.

Such sorting methods are clear and easy to understand and standardise, and can help farmers to position their farm among others and to make them more aware of the conditions in which their animals are living.

In these methods, each partial rank (i.e. each measure) has the same importance in the final overall rank. For instance, in the method proposed by Whay *et al.* (2003), hind limb dirtiness was as important as pneumonia. However, it is possible to assign a weighting to measures, for instance by counting the rank obtained for a particular measure twice in the calculation of the final score if this measure is considered to be twice as important as each of the others.

The main limit of this method is that the overall ranking of a farm depends on the population observed: a better ranking obtained by one farm over another results not only from their relative positions in the rankings for the different measures but also from their relative positions with respect to the other farms examined (Bouyssou *et al.*, 2000, p. 16; see Table 1 for a numerical example). The Borda method should be used only in very specific situations in which the set of farms is perfectly well-defined and stable over time. This will not generally be the case in a voluntary labelling system.

Sum (or mean) of scores

While a sum of ranks allows comparisons only within a given sample of observations, a sum of scores produces absolute values (i.e. independent of the sample observed). Such methods are widely developed in the context of animal welfare (Bartussek, 1999; Keeling and Svedberg, 1999; Horning, 2001; Scott *et al.*, 2001; Bracke *et al.*, 2002).

A common framework can be seen for all these methods. First, raw data expressed in very different units (e.g. kg for body weight, % of animals affected by mastitis and *score* for body condition) are converted into a 'partial welfare score' on a numerical value scale with a meaning that is common to all the measures (i.e. a commensurable scale). Second, weightings are assigned to the values obtained for the different measures to account for the impact of each measure on animal welfare. These weightings are explicit in Scott *et al.* (2001) and Bracke *et al.* (2002). They are implicit in the Tier Gerechtheits Index (TGI) defined by Bartussek (1999) or Sundrum and Rubelowski (2001) where they correspond to the size of the value scale: e.g. in TGI35L, a 0 to 3 scale is used for space allowance and a −0.5 to 0.5 scale is used for cleanliness. Third, an overall score is calculated as the weighted sum (or weighted mean) of the partial scores obtained for each measure.

Some authors, however, do not use weightings. For example, El Balaa and Marie (2004) proposed a method to assess the welfare of small ruminants where a total of

https://doi.org/10.1017/S1751731107000535 Published online by Cambridge University Press

Table 1 *Illustration of how a sum of ranks depends on the sample of units compared with each other*

	Scores		Partial ranks			
Farms	Measure 1	Measure 2	Measure 1	Measure 2	Sum of ranks	Comparison between A and B
First set of farms						
A	10	15	3	1	**4**	A is preferred to B
B	16	8	1	4	**5**	
C1	6	9	4	3	7	
D1	12	13	2	2	4	
Second set of farms						
A	10	15	3	1	**4**	B is preferred to A
B	16	8	1	2	**3**	
C2	9	5	4	4	8	
D2	13	7	2	3	5	

Two farms A and B are compared on the basis of two distinct measures. At the top of the table, where farms A and B are compared with farms C1 and D1, farm A ranks higher than B. At the bottom of the table, where farms A and B are compared with farms C2 and D2, farm B ranks higher than A

Table 2 *(a and b) Illustration of how differences in scaling can affect welfare outcomes when a sum of scores is calculated*

(a) 'Lying-down movements' is expressed on a three-level scale: normal movement after only one intention (0); several intentions before a normal lying movement (1); interrupted movement (one foreleg is bent then the animal gets up) (2). 'Reactions to human' is expressed on a five-level scale: the animal approaches the experimenter (0); it does not move (0.5); it stops eating (1); it makes one step back (1.5); it makes several steps back (2). The welfare of A and B is concluded to be equal

	Scale	Animal A	Animal B	Comparison between A and B
Lying-down movements	0/1/2	2	1	B is better than A
Response to human approach	0/0.5/1/1.5/2	0	1	A is better than B
Sum of scores		2	2	The welfare of A and B are similar

(b) 'Lying down movements' is expressed on a three-level scale (the same as for Table 2a), and 'reactions to human' is expressed on a three-level scale: the animal approaches (0); it does not move (1); it stops eating or moves back (2). The welfare of A is considered to be higher than that of B

	Scale	Animal A	Animal B	Comparison between A and B
Lying-down movements	0/1/2	2	1	B is better than A
Response to human approach	0/1/2	0	2	A is better than B
Sum of scores		2	3	The welfare of A is higher than the welfare of B

Two animals A and B are compared on the basis of lying movements (continuous lying movement is considered to be better for animal welfare) and responses to humans (not being afraid of humans being considered better). Animal A displays an interrupted lying down movement and approaches the experimenter while Animal B had no interruption of lying and stops eating when the experimenter comes near. The preference between A and B is assigned according to the sum of partial scores obtained for these two measures

43 measures is used to check compliance with the five freedoms defined by FAWC (1992). For each freedom, they calculated the mean of the scores obtained for the measures related to that freedom, giving the same importance to all measures.

Weighted sums (or means) of scores are very popular and are easily understood by non-scientists, at least in their general principles. Partial scores can be used to point out strong *v.* weak points of each farm assessed and so can help farmers to choose ways to improve the welfare of their animals. The overall score allows comparisons between animal units while an absolute judgement of a farm, independently of the others, can still be made. However, summing of scores suffers from several limitations. These are presented below.

First, welfare scores must be cardinal values, i.e. data assessed on interval or ratio scales (Bouyssou *et al.*, 2000, pp. 47–48), to be appropriate for the calculation of a weighted sum. However, it is often assumed, rather than shown, that the intervals between levels of the scales used are equivalent. Uncertainty about intervals between ordinal scales can lead to confusing results. This point is illustrated with a numerical example in Table 2. Ideally, the scales for different welfare measures should be converted into a unified scale (e.g. a scale between 0, worst for welfare, and 1, best for welfare) before (weighted) sums (or means) of scores are calculated.

Second, sums of scores allow full compensation. Spoolder *et al.* (2003) pointed out that when a sum of scores is calculated, a serious welfare disadvantage can be

https://doi.org/10.1017/S1751731107000535 Published online by Cambridge University Press

compensated for by a number of minor advantages, even though this may not always be legitimate. Full compensation between welfare aspects may not be desirable. This point will be further discussed in part 2 of the present dissertation on overall assessment of animal welfare (Botreau *et al*., 2007). One way to limit compensation could be to specify minimal requirements below which the animal unit is discarded (Perny, 1998). Bracke *et al*. (2002) introduced very high weighting factors for some levels of partial scores ($\pm 10\,000$ compared with ± 1, ± 2, ± 3, ± 5 used in most cases in the model). These very high weightings act as minimum requirements, limiting the possibility of compensating for a serious welfare problem with a number of small benefits.

In addition, when compensation between measures is legitimate, we can accept that a small increase in an average score for one measure (e.g. going from 15% to 20% of mastitis in a dairy herd) may be compensated for by a small decrease in an average score for another measure (e.g. going from 20% to 15% in lameness). However, compensation may not be acceptable at other points on the scale. For instance, an increase from 0% to 5% of mastitis may not be compensated for by a decrease from 20% to 15% of lameness because 0 mastitis means that there is no risk to spread mastitis from one animal to another, whereas at 5% there is some risk. Hence, when using sums of scores, attempts should be made to ensure that trade-offs are constant.

Third, sums (or means) of scores do not favour situations of compromise (Bouyssou *et al*., 2000, pp. 43–45). A farm that obtains average scores for all measures (i.e. a homogeneous profile) could obtain the same overall score as a farm with very low scores on some measures and very high ones on others (i.e. a heterogeneous profile). However, from an animal's point of view, it may be preferable to live on a farm with moderate results on all measures than to be subjected to very poor environmental conditions irrespective of the other aspects. This point can be illustrated looking at three animals compared on the basis of their responses to human approach and their lying-down movement (the scales are the same as those used in Table 2b). In this example, Animal *X* and Animal *Y* have contrasting partial scores:

- *X* is not at all afraid of humans – i.e. scores 0 on measure 1 – but displays serious difficulties lying down – i.e. scores 2 on measure 2;
- *Y* is very afraid of humans – i.e. scores 2 on measure 1 – but has no problem at all lying down – i.e. scores 0 on measure 2;
- *Z* is moderately good for both measures – slightly disturbed by the presence of humans – i.e. scores 1 on measure 1 – and hesitates before lying down – i.e. scores 1 on measure 2.

The sum of scores equals 2 for *X*, *Y* and *Z*. However, a best compromise might be to have a mild response to human approach and to have only one intention before lying, rather than to approach humans but to have interrupted lying down movements. But whatever the weightings used to sum the two scores, *Z* will never be strictly preferred to *X* and *Y*: let w_1 and w_2 be the weights associated with measure 1 and measure 2;

- to consider *Z* in a better state than *X*, w_1 and w_2 shall be chosen so that $w_1 + w_2 < 2w_2$, i.e. $w_1 < w_2$;
- to consider *Z* in a better state than *Y*, w_1 and w_2 shall be chosen so that $w_1 + w_2 < 2w_1$, i.e. $w_2 < w_1$.

These two conditions '$w_1 < w_2$' and '$w_2 < w_1$' are obviously not consistent!

In conclusion, the sum of scores is very intuitive. It requires a number of conditions (cardinal data and constant trade-offs) that are difficult to check in practice. In addition, it may allow compensation where compensation should be restricted, and cannot favour compromises. Hence, it is hazardous to use such a method to produce an overall assessment based on non-compensatory aspects. However, it may be an appropriate method for compounding subsets of welfare measures related to the same welfare aspect, where compensation between measures can be considered legitimate (e.g. summing the occurrences of major diseases to obtain an overall view of the health status of a herd, as in Mounier *et al*., 2006). In the same vein, multivariate analyses could be used to construct welfare indices. The first components of principal component analyses (PCA) are linear combinations of variables, i.e. weighted sums with the weight assigned to a variable depending on its distribution in relation to other variables and thus on its potential to describe variations in the population studied (Lebart and Fenelon, 1975). PCA have been successfully used on subsets of variables related to the same welfare issue. For instance, Veissier and Capdeville (2004) considered three main criteria describing the comfort of cows cubicles: difficulties in lying down and getting up movements, injuries, cleanliness. For each criterion, several measures were defined, a PCA was run on the data obtained from a survey on 70 farms, and the first component of PCA was considered as a summary variable to compare several cubicle designs. Such a use of PCA appears relevant in that context because within a criterion the variables can be considered as different ways of looking at the same problem (e.g. difficulties in lying down are generally accompanied by difficulties in getting up). However, setting the weights for separate criteria on the basis of a PCA seems to us inappropriate because there is no functional links between criteria. Furthermore, the limitations of weighted sums highlighted above would fully apply to PCA used to combine several welfare criteria.

Use of *ad hoc* rules limiting compensation

Capdeville and Veissier (2001) proposed using *ad hoc* rules to make an overall assessment of animal welfare for loose-housed dairy cows. Their rules were based on expert

https://doi.org/10.1017/S1751731107000535 Published online by Cambridge University Press

Table 3 *Currently proposed methods for the overall assessment of animal welfare at farm level, with their advantages and limitations, and their potential use*

Method	Advantages	Limitations	Potential uses
Non-formal aggregation, an expert's opinion is produced on the examination of a report gathering all the data collected	• Based only on the raw data collected on farm, i.e. with no calculation	• Lack of transparency • Impossible to standardise for routine use by non-experts • A large amount of data is difficult to compound by the expert	• Perfect for the assessment of a few animal units or housing systems/equipments. This method should be limited to an analysis on a case-by-case basis.
'Checklist', each measure is compared with a minimal requirement	• Clear and simple • Easy to standardise • Allows correspondence to a standard/norm to be checked	• Yields an 'all-or-nothing' response • All measures have the same importance • Does not allow comparisons between farms	• Routine use, to check that all requirements are fulfilled, leading to a yes/no answer. This method can be used to check the respect for the current laws or within certification schemes based on the strict fulfilment of a series of specifications.
Sum of ranks, ranks obtained on several measures by a farm within a given set are summed	• Clear and simple • Easy to standardise • Allows ranking of farms	• All measures have the same importance • Allows comparisons between farms only within a given set of farms	• Routine use, to rank animal units belonging to a pre-set group of farms.
Weighted sum of scores, scores obtained on several measures by a farm are summed	• Relatively intuitive • Allows an absolute score for any farm to be produced	• Compensations fully authorised between measures • Does not favour situations of compromise	• Routine use, leads to an answer more sensitive than a mere yes/no answer. This method can be used to compare farms and/or to implement certification or labelling schemes where interactions between welfare items are permitted

opinion and were designed to limit compensation between measures. A set of 16 animal needs was defined, derived from the five freedoms (FAWC, 1992). Forty-nine measures were recorded on-farm to assess how these needs were met (essentially animal-based measures such as lying movements, lameness, injuries and agonistic social behaviour). As in the methods based on sum of scores (see above), a value scale was used: 'A' corresponded to a very high level of welfare; 'B', a moderately high level of welfare; 'C', a moderately low level of welfare; and 'D', a very low level of welfare. The difference from the previous methods lies in the rules used to aggregate the information. For each possible combination of partial scores, an overall value was decided on and this is generally lower than the mean of partial scores. For instance, a score 'A' on one measure with a score 'B' on another measure results in a 'B' for the overall value for the two measures taken together. Capdeville and Veissier (2001) defined three types of rules to aggregate information according to the importance of measures for the welfare of animals. The rules were stricter, i.e. allowed less compensation, or even forbade it, for measures they considered especially important (e.g. severe injuries).

In this method, and also in some methods using weighted sums, the aggregation process is performed in different stages: first a few measures are grouped together, then groups of measures are aggregated, etc. For instance, 'lying-down and standing-up movements' and 'resting posture' are considered together to assess resting behaviour before being included in the need for physical comfort. This hierarchical process limits problems of overweighting items that are assessed in numerous measures. Finally, a score is given to the farm for each basic need and then for the five freedoms together.

The *ad hoc* rules defined by Capdeville and Veissier (2001), however, could group only a few items together (maximum = 4). This required making many small sets of measures and then making many sequential aggregations. Consequently, the grouping may not always have a biological meaning and the entire method is rather opaque and difficult to explain.

In conclusion, this method limits compensation between measures and can favour compromise situations. However, the opacity of the method makes it difficult to appraise its legitimacy and to transpose it to other animal types.

https://doi.org/10.1017/S1751731107000535 Published online by Cambridge University Press

Discussion

The advantages and limits of each category of methods for aggregating welfare measures analysed in this paper are summarised in Table 3, together with specifications on the context in which they are most appropriate. The choice of a method may in part depend on the objective for the use of a welfare assessment. Comparisons with thresholds are better suited to check compliance with requirements, whereas weighted sums can help to compare farms or farming systems. However, we feel that none of the existing methods proposed to produce an overall assessment of animal welfare is yet fully satisfactory to be used within a certification scheme, which includes several welfare classes (i.e. more sensitive than a mere yes/no classification). The most common shortcomings that limits application in a certification scheme are:

(i) allowing too much compensation between welfare aspects (e.g. in sum of ranks or sum of scores), and this may result in overshadowing a serious welfare problem by many small welfare improvements;
(ii) assuming all measures have the same importance (e.g. in definition of minimal requirements for measures or sum of ranks), and this does not help focussed improvements to be targeted on the most serious problems;
(iii) difficulty in creating a formal structure for routine use (e.g. in non-formal aggregation or use of *ad hoc* rules), which reduces the potential for using the method on a large scale and may result in difficulties in communicating the results;
(iv) providing assessments that are meaningful only for a given data set (sum of ranks), which may lead to competition between farmers rather than true welfare improvements.

Some authors propose systems that include different methods for the aggregation of measures. For example, Keeling and Svedberg (1999) proposed the use of 36 welfare measures on poultry units. For 12 measures, they defined thresholds corresponding to minimum legislative requirements. When these requirements were satisfied, 24 other measures were assessed on the farm and were further compounded using a weighted sum to produce an overall estimation of the welfare of animals.

Such a mixture of aggregation methods have already been used in other contexts, for instance in the European New Car Assessment Programme (EURONCAP, 2004). New cars are assessed for the level of protection displayed for adult occupants, children and pedestrians. The child protection assessment leads to a rating from 0 to 5 stars, according to a sum of partial scores obtained for each issue of child protection. Several methods are used to attribute partial scores: for some issues a checklist with minimum requirements is used, for some other issues decision trees are used.

This use of 'mixed methods' is probably to be encouraged so that a range of specific features linked to welfare assessment can best be taken into account. For instance, compensations may be allowed between measures that represent different viewpoints on the same problem (e.g. jointly considering difficulties both in lying down, and in getting up, to assess comfort around resting). However, it seems hazardous to consider that, for example, good health can fully compensate for the reduced ability to express natural behaviours. When several welfare dimensions are to be combined, non-compensatory methods such as the definition of minimal requirements seem more appropriate. Hence, different methods may be used at different stages in the construction of the overall assessment depending on the constraints imposed on the aggregation process. Therefore, the constraints imposed on the aggregation of welfare measures need to be clearly identified at a number of points in the aggregation process to generate meaningful results. These constraints are further described and analysed in Part 2 of the present dissertation on the overall assessment of animal welfare (Botreau *et al.*, 2007).

Acknowledgements

The present study is part of the Welfare Quality® research project, which has been co-financed by the European Commission, within the sixth Framework Programme, contract no. FOOD-CT-2004-506508. The text represents the authors' views and does not necessarily represent a position of the Commission who will not be liable for the use made of such information.

References

Algers B, Ekstrand C, Geismar J, Gunnarsson S, Odén K, Onila M and Svedberg J 1995. Utvärden av OLI voletage inhysningssystem för värphöns. I. enlighet med SJV:s program för förprövning av ny teknik. Swedish University of Agricultural Sciences, Skara, Sweden.

Algers B, Broom D, Canali E, Hartung J, Smulders F, Van Rennen CG and Veissier I 2006. Scientific opinion on the risk of poor welfare in intensive calf farming systems: an update of the Scientific Veterinary Committee Report on the Welfare of calves. The EFSA Journal 366, 1–144.

Barnett JL and Hemsworth PH 1990. The validity of physiological and behavioural measures of animal welfare. Applied Animal Behaviour Science 25, 177–187.

Bartussek H 1999. A review of the animal needs index (ANI) for the assessment of animals' well-being in the housing systems for Austrian proprietary products and legislation. Livestock Production Science 61, 179–192.

Bell DE, Raiffa H and Tversky A 1988. Decision making: descriptive, normative and prescriptive interactions. Cambridge University Press, Cambridge.

Bock B and van Leeuwen F 2005 Socio-political and market developments of animal welfare schemes. In Farm animal welfare concerns, consumers, retailers and producers – WelfareQuality® reports no. 1 (ed. J Roex and M Miele), pp. 115–167. Cardiff University.

Botreau R, Bracke MBM, Perny P, Butterworth A, Capdeville J, van Reenen CG and Veissier I 2007. Aggregation of measures to produce an overall assessment of animal welfare. Part 2: analysis of constraints. Animal 1.

Bouyssou D, Marchant T, Pirlot M, Perny P, Tsoukias A and Vincke P 2000. Evaluation and decision models – a critical perspective. Kluwer Academic Publishers, Dordrecht.

Bracke MBM, Spruijt BM, Metz JHM and Schouten WGP 2002. Decision support system for overall welfare assessment in pregnant sows. A model structure and weighting procedure. Journal of Animal Science 80, 1819–1834.

Capdeville J and Veissier I 2001. A method of assessing welfare in loose housed dairy cows at farm level, focusing on animal observations. Acta Agriculturae Scandinavica, Section A, Animal Science Supplementum 30, 62–68.

https://doi.org/10.1017/S1751731107000535 Published online by Cambridge University Press

Dawkins MS 1980. Animal suffering: the science of animal welfare. Chapman & Hall Ltd, London.

Dawkins MS 1990. From an animal's point of view: motivation, fitness, and animal welfare. Psychological Science 13, 1–61.

El Balaa R, Marie M 2004. Evaluation du bien-être animal dans les élevages de petits ruminants. Animal welfare evaluation in small ruminant husbandry. Proceedings of the 11th Rencontres Recherches Ruminants, p. 210. Institut de l'Elevage – INRA, Paris.

Ekstrand C, Odén K, Gunnarsson S, Onila M, Algers B and Svedberg J 1997. Utvärden av OLI voletage inhysningssystem för värphöns. I. enlighet med SJV:s program för förprövning av ny teknik. Swedish University of Agricultural Sciences, Skara, Sweden.

EUREPGAP nd. EUREPGAP – the global partnership for safe and sustainable agriculture. Retrieved June 4, 2007, from http://eurep.org/.

European New Car Assessment Programme, 2004. Child protection assessment protocol. Retrieved June 4, 2007, from http://www.euroncap.com/Downloads/31f3c2e9-9206-4d74-9e2f-cc9908430621/Child-Assessment-Protocol-ver-1-0c.pdf.aspx.

Farm Animal Welfare Council 1992. FAWC updates the five freedoms. The Veterinary Record 17, 357.

Fraser D 1993. Assessing animal well-being: common sense, uncommon science. Proceedings of the food animal well-being conference proceedings and deliberations (ed. USDA and Purdue University Office of Agricultural Research Programs), pp. 37–54. Purdue University Office of Agricultural Research Programs, West Lafayette, Indiana.

Fraser D 1995. Science, values and animal welfare: exploring the 'inextricable connection'. Animal Welfare 4, 103–117.

Fraser D 2003. Assessing animal welfare at the farm and group level: the interplay of science and values. Animal Welfare 12, 433–443.

French S 1984. Fuzzy decision analysis: some criticism. TIMS/Studies in the Management Sciences 20, 29–44.

Friend TH 1980. Stress: what is it and how can it be quantified? International Journal for the Study of Animal Problems 1, 366–374.

Geers R, Petersen B, Huysmans K, Knura-Deszczka S, De Becker M, Gymnich S, Henot D, Hiss S and Sauerwein H 2003. On-farm monitoring of pig welfare by assessment of housing, management, health records and plasma haptoglobin. Animal Welfare 12, 643–647.

Hegelund L, Sørensen JT and Johansen NF 2003. Developing a welfare assessment system for use in commercial organic egg production. Animal Welfare 12, 649–653.

Horning B 2001. The assessment of housing conditions of dairy cows in littered loose housing systems using three scoring methods. Acta Agriculturæ Scandinavica, Section A, Animal Science Supplementum 30, 42–47.

Humane Farm Animal Care, 2003. Certified humane raised and handled. Retrieved June 4, 2007, from http://certifiedhumane.com.

Hurnik JF 1990. World's poultry science association invited lecture: animal welfare: ethical aspects and practical considerations. Poultry Science 69, 1827–1834.

Huxley JN, Burke J, Roderick S, Main DCJ and Whay HR 2004. Animal welfare assessment benchmarking as a tool for health and welfare planning in organic dairy herds. Veterinary Record (The) 155, 237–239.

Johnsen PF, Johannesson T and Sandoe P 2001. Assessment of farm animal welfare at herd level: many goals, many methods. Acta Agriculturæ Scandinavica, Section A, Animal Science Supplementum 30, 26–33.

Keeling L, Svedberg J 1999. Legislation banning conventional battery cages in Sweden and a subsequent phase-out programme. Proceedings of the Congress 'Regulation of Animal Production in Europe' (ed. M Kunisch and H Eckel), pp. 73–78.

Lacroix R, Huijbers J, Tiemessen R, Lefebvre D, Marchand D and Wade KM 1997. Fuzzy set-based analytical tools for dairy herd improvement. Applied Engineering in Agriculture 14, 79–85.

Lebart L and Fenelon JP 1975. Statistique et informatique appliqués. Editions Dunod, Paris.

Main DCJ, Webster F and Green LE 2001. Animal welfare assessment in farm assurance schemes. Acta Agriculturæ Scandinavica, Section A, Animal Science Supplementum 30, 108–113.

Main DCJ, Kent JP, Wemelsfelder F, Ofner E and Tuyttens FAM 2003. Applications for methods of on-farm welfare assessment. Animal Welfare 12, 523–528.

Mounier L, Colson S, Roux M, Dubroeucq H, Boissy A, Ingrand S and Veissier I 2006. Links between specialization in the finishing of bulls, mixing, farmers' attitudes towards animals and the production of finishing bulls: a survey on French farms. Animal Science 82, 561–568.

Perny P 1998. Multicriteria filtering methods based on concordance and non-discordance principles. Annals of Operations Research 80, 137–165.

Piñeiro C, Morales M, Piñeiro M, Ruiz de la Torre JL, Mateos GG, Manteca X. 2005. Psychological stress caused by changes in feeding management did not affect the concentration of acute phase proteins in pigs. Proceedings of the Fifth International Colloquium on Animal Acute Phase Proteins.

Roy B 1993. Decision science or decision-aid science? European Journal of Operation Research 66, 184–204.

Royal Society for the Prevention of Cruelty to Animals undated. Flow chart for the Freedom Food scheme. Retrieved June 4, 2007, from http://www.rspca.org.uk/servlet/Satellite?pagename=RSPCA/RSPCARedirect&pg=Producer section&marker=1&articleId=1121758148522.

Rushen J 1986. The validity of behavioural measures of aversion: a review. Applied Animal Behaviour Science 16, 309–323.

Rutter SM, 1998. Assessing the welfare of intensive and extensive livestock. Proceedings of the Workshop Pasture Ecology and Animal Intake, pp. 1–9.

Scientific Committee on Animal Health and Animal Welfare of the European Commission undated. Reports from the Scientific Committee on Animal Health and Animal Welfare of the European Commission. Retrieved June 4, 2007, from http://ec.europa.eu/food/fs/sc/scah/outcome_en.html.

Scott EM, Nolan AM and Fitzpatrick JL 2001. Conceptual and methodological issues related to welfare assessment: a framework for measurement. Acta Agriculturæ Scandinavica, Section A, Animal Science Supplementum 30, 5–10.

Sørensen JT, Sandoe P and Halberg N 2001. Animal welfare as one among several values to be considered at farm level: the idea of an ethical account for livestock farming. Acta Agriculturæ Scandinavica, Section A, Animal Science Supplementum 30, 11–16.

Spoolder H, De Rosa G, Horning B, Waiblinger S and Wemelsfelder F 2003. Integrating parameters to assess on-farm welfare. Animal Welfare 12, 529–534.

Sundrum A and Rubelowski I 2001. The meaningfulness of design criteria in relation to the mortality of fattening bulls. Acta Agriculturæ Scandinavica, Section A, Animal Science Supplementum 30, 48–52.

Veissier I, Capdeville J and Delval E 2004. Cubicle housing systems for cattle: comfort of dairy cows depends on cubicle adjustment. Journal of Animal Science 82, 3321–3337.

Von Borell E, Bockisch FJ, Buscher W, Hoy S, Krieter J, Muller C, Parvizi N, Richter T, Rudovsky A, Sundrum A and Van Den Weghe H 2001. Critical control points for on-farm assessment of pig housing. Livestock Production Science 72, 177–184.

Webster J 1997. Applied ethology: what use is it to animal welfare? Advances in ethology 32Supplements to Ethology, 10.

Webster J 2005. The assessment and implementation of animal welfare: theory into practice. Revue Scientifique Et Technique-Office International Des Epizooties 24, 723–734.

Wechsler B 2001. Pretesting of mass-produced farm animal housing systems in Switzerland 20 years of experience. Proceedings of the international symposium of the second technical section of CIGR on animal welfare considerations in livestock housing systems (ed. Polish Committee of Agricultural Engineering), pp. 55–67. Poligmar, Zielona Góra, Poland.

Wechsler B 2003. Testing of mass-produced farm animal housing systems with regard to animal welfare. Proceedings of the 54th annual meeting of the European Association for Animal Production (ed. Y van der Honing), p. 125. Wageningen Pers, Wageningen.

Wechsler B 2005. An authorisation procedure for mass-produced farm animal housing systems with regard to animal welfare. Livestock Production Science 94, 71–74.

Whay HR, Main DCJ, Green LE and Webster AJF 2003. An animal-based welfare assessment of group-housed calves on UK dairy farms. Animal Welfare 12, 611–617.

Winckler C, Capdeville J, Gebresenbet G, Horning B, Roiha U, Tosi M and Waiblinger S 2003. Selection of parameters for on-farm welfare-assessment protocols in cattle and buffalo. Animal Welfare 12, 619–624.

https://doi.org/10.1017/S1751731107000535 Published online by Cambridge University Press